CLAIMS

What is claimed is:

1. A system for summarizing audio information, comprising:

an analyzer to convert audio into frames;

a fingerprinting component to convert the frames into fingerprints, each fingerprint based in part on a plurality of frames;

a similarity detector to compute similarities between fingerprints;

a heuristic module to generate a thumbnail of the audio file, based in part on the similarity between fingerprints.

2. The system of claim 1, the heuristic module comprising at least one of an energy component and a flatness component in order to help determine a suitable segment of audio for the thumbnail.

3. The system of claim 2, the heuristic module is employed to automatically select voiced choruses over instrumental portions.

4. The system of claim 2, the energy component and the flatness component are employed when the fingerprints do not result in finding a suitable chorus.

5. The system of claim 1, further comprising a component to remove silence at the beginning and end of an audio clip *via* an energy-based threshold.

6. The system of claim 1, the fingerprint component further comprising a normalization component, such that an average Euclidean distance from the each fingerprint to other fingerprints for an audio clip is one.

7. The system of claim 1, the analyzer computes a set of spectral magnitudes for an audio frame.

8. The system of claim 7, for each frame, a mean, normalized energy E is computed by dividing a mean energy per frequency component within the frame by the average of that quantity over frames in an audio file.

9. The system of claim 8, further comprising a component that selects a middle portion of an audio file to mitigate quiet introduction and fades appearing in the audio file.

10. The system of claim 2, the flatness component employs a number that is added to spectral magnitudes for each frequency component, to mitigate numerical problems when determining logs.

11. The system of claim 10, the flatness component includes a frame-quantity computed as a log normalized geometric mean of the spectral magnitudes.

12. The system of claim 11, the normalization is performed by subtracting a per-frame log arithmetic mean of a per-frame magnitudes from the geometric mean.

13. The system of claim 1, the similarity detector comprising a clustering function, the clustering function producing clusters of similar fingerprints.

14. The system of claim 13, the clustering function further producing sets of clusters.

15. The system of claim 14, further comprising a fingerprint F1 or an identifying index related to F1 that is added to a cluster containing fingerprint F2 in the cluster set if F1 and F2 satisfy at least two conditions: with respect to a first condition, a normalized Euclidean distance from F1 to F2 is below a first threshold, and with respect to a second condition, a temporal gap in an audio between where F1 is computed and where F2 is computed is above a second threshold.

16. A computer readable medium having computer readable instructions stored thereon for implementing the system of claim 1.

17. An automatic thumbnail generator, comprising:

means for converting an audio file into frames;

means for fingerprinting the audio file, producing fingerprints based in part on a plurality of frames; and

means for determining an audio thumbnail based in part on the fingerprints.

18. A method to generate audio thumbnails, comprising:

generating a plurality of audio fingerprints, each audio fingerprint based in part on a plurality of audio frames;

clustering the plurality of fingerprints into fingerprint clusters; and

creating a thumbnail based in part on the fingerprint clusters.

19. The method of claim 18, the clustering further producing one or more cluster sets, each cluster set comprising fingerprint clusters.

20. The method of claim 19, the clustering further comprising determining whether a cluster set has three or more fingerprint clusters.

21. The method of claim 18, the clustering based in part on a threshold, the threshold chosen adaptively for an audio file and used to help determine if two fingerprints belong to the same cluster set.

22. The method of claim 18, the clustering operating by considering one fingerprint at a time.

23. The method of claim 18, further comprising determining a parameter (D) describing how evenly spread clusters are, temporally, throughout an audio file.

24. The method of claim 23, wherein a measure of temporal spread is applied to the clusters in a given cluster set.

25. The method of claim 24, (D) is measured as follows:

normalizing a song to have duration of 1;

setting a time position of an $i'th$ cluster be t_i;

defining $t_0 \square 0$ and $t_{N+1} \square 1$; and

computed as $\frac{(N+1)}{N}\left(1-\sum_{i=1}^{N+1}(t_i-t_{i-1})^2\right)$ where N is a number of clusters in a cluster set.

26. The method of claim 25, further comprising determining an offset and scaling factor so that (D) takes a maximum value of 1 and minimum value of 0, for any N.

27. The method of claim 25, further comprising determining a mean spectral quality for fingerprints in a set.

28. The method of claim 27, wherein a mean spectral flatness for a set, and a parameter D, are combined to determine a best cluster set from among a plurality of cluster sets.

29. The method of claim 28, the mean spectral flatness and parameter D are combined into a single parameter associated with each cluster set, such that the set with the external value of the parameter is selected to be the best set.

30. The method of claim 29, when the best cluster set is selected, a best fingerprint within the cluster set is determined as the fingerprint in which surrounding audio, of duration about equal to a duration of an audio thumbnail, has maximum spectral energy or flatness.

31. The method of claim 18, the creating further comprising determining a cluster by determining a longest section of audio within an audio file that repeats in the audio file.

32. The method of claim 18, the creating further comprising at least one of:

rejecting clusters that are close to a beginning or end of a song;

rejecting clusters for which energy falls below a threshold for any fingerprint in a predetermined window; and

selecting a fingerprint having a highest average spectral flatness measure in the predetermined window.

33. The method of claim 18, the creating further comprising generating a thumbnail by specifying time offsets in an audio file.

34. The method of claim 18, the creating further comprising automatically fading a beginning or an end of an audio thumbnail.

35. The method of claim 18, the generating further comprising processing an audio file in at least two layers, where the output of a first layer is based on a log spectrum computed over a small window and a second layer operates on a vector computed by aggregating vectors produced by the first layer.

36. The method of claim 35, further comprising providing a wider temporal window in a subsequent layer than a proceeding layer.

37. The method of claim 36, further comprising employing at least one of the layers to compensate for time misalignment.